This document is published in:

Andre Ponce de Leon F. de Carvalho et al. (eds.) (2010). *Distributed Computing and Artificial Intelligence*: *7th International Symposium*. (Advances in Intelligent and Soft Computing, 79) Springer, 325-332.
DOI: http://dx.doi.org/10.1007/978-3-642-14883-5_42

Interactive Video Annotation Tool

Miguel A. Serrano, Jesús García, Miguel A. Patricio, and José M. Molina

GIAA, Carlos III University, Spain
e-mail: {miguel.serrano,jesus.garcia}@uc3m.es,
{miguelangel.patricio,josemanuel.molina}@uc3m.es

Abstract. Increasingly computer vision discipline needs annotated video databases to realize assessment tasks. Manually providing ground truth data to multimedia resources is a very expensive work in terms of effort, time and economic resources. Automatic and semi-automatic video annotation and labeling is the faster and more economic way to get ground truth for quite large video collections. In this paper, we describe a new automatic and supervised video annotation tool. Annotation tool is a modified version of ViPER-GT tool. ViPER-GT standard version allows manually editing and reviewing video metadata to generate assessment data. Automatic annotation capability is possible thanks to an incorporated tracking system which can deal the visual data association problem in real time. The research aim is offer a system which enables spends less time doing valid assessment models.

Keywords: Ground Truth, Tracking, Automatic Annotation.

1 Introduction

Over the last years, the amount of multimedia resources has grown due to the popularization of Web 2.0. This information is unstructured and poorly organized, being very hard to browse and retrieve it. In order to overcome this limitation, it is necessary to semantically label and organize all this data.

Annotation is a process which provides visual metadata superimposed over resources without modifying the analyzed element. Manual annotation is an unfeasible task for large video collections however automatic video annotation systems can automatically add metadata through computer vision techniques.

Normally, new automatic systems based on computer vision techniques try to improve their benefits comparing with the current state of art. To demonstrate these improvements in a scientific way it is only possible through assessment models.

Therefore annotation and label systems need tools to assess the performance of their techniques. Such evaluation is often carried out by comparing results obtained from a given algorithm against ground truth - a set of results determined a

priori to be correct [4]. More specifically, in the video evaluation scope, generate ground truth annotations for large scale video collections has involved huge amount of effort. Traditional techniques don't work because a temporal dimension generated from video frame sequence is added to the images spatial dimensions.

This paper presents a new supervised automatic annotation tool. The basic infrastructure is a tracking system integrated to an extended version of the ViPER-GT annotation tool.

Perform tracking tasks during the video analysis, facilitates the automatic annotation feature. In tracking low level tasks, such as segmentation or trajectory analysis, the tool detects, label and annotates tracks. In addition, user may manually create tracks or modify the location, size and trajectory if an error occurs.

Moreover, with this tool high level semantic tasks at scene and object level can be developed. Semantic annotations are done manually trough the ViPER-GT tool interface. All these actions can be done in real time because the system is adaptable to the changes.

The paper is organized as follows. In Section 2 annotation and assess fields are studied briefly; Section 3 the annotation tool overall architecture is presented; Section 4 shows the experimental results; Section 5 explains the conclusions obtained and the future work.

2 Brief Summary about Annotation

For years researchers in annotation have been worked in two different ways. Initial approaches focused on low level visual descriptors such as texture, shape… After, researches turned to knowledge approaches, which try to extract semantic descriptions with the goal of save the semantic gap.

Low level descriptors approaches imitate the way users assess visual similarity [5] and don't try to extract directly semantic assertions from visual content. The main feature of this kind of methods is the capability to find patterns from frame/image features. These techniques are based on machine learning methods. Most used techniques in this field are Hidden Markov Models and Neural Networks.

On the other hand, knowledge-based approaches uses a higher abstraction level when annotate content. To that end, make use of "a priori" knowledge such as models, rules… These approaches, normally allow realizing inference operations between the elements and the spatial relationships. New hidden domain knowledge results of these operations.

The actual trend is to blend both approaches, extracting relevant semantic elements from videos by combining several low-level descriptors [7]. In this context, one of the keys is the MPEG-7 standard. MPEG-7 represents audiovisual information and allows content descriptions. For instance, MPEG-7 Visual Part support low level features such as color, texture, shape or motion. There are also MPEG-7 Multimedia Description Schemes which support spatial relations between detected segments.

Nowadays general purpose approaches don't exist due to the knowledge dependency with the specific context domain. High level semantic applications are based on specific context such as sports, movies, security…

Inside evaluation field there are some resources such as databases and semi-automatic annotation tools. Large scale general purpose databases are scarce. Examples of publicly available sets of databases are: NIST TRECVID databases which contain several hours of publicly available *ground truth*, from over 1000 visual concept categories, PETS datasets [18] which actually include outdoor people and vehicle tracking, indoor people tracking, annotated hand posture classification data..., the Surveillance Performance Evaluation Initiative (SPEVI) [12] which includes an audiovisual people dataset, a single face dataset and a multiple face dataset all of them made with the Video Performance Evaluation Resource (ViPER), on the other hand as single dataset we can list, cVSG [13], OTCBVS [14], VISOR [15], ETISEO [16], CANDELA [17], etc.

We can also find notable tools to create new *ground truths*, the IBM MPEG-7 Annotation Tool, for example, provides a rich user interface that displays the video, semi-automatically detects shot boundaries and selects key frames, automatically propagates prior shot labels, presents a hierarchy of 133 suggested visual concept labels but also accepts new user-created visual concept labels [8]. ViPER Ground Truth (ViPER-GT) is another interesting video annotation tool, which may be used as a viewer of algorithmically generated markup, a tool for assisting performance evaluation of such markup and more. ViPER Performance Evaluation tool (ViPER-PE) complete the capabilities of ViPER-GT providing the ability to compare result data with ground truth tools for solving the evaluation problem [11].

3 Overall Architecture

Annotation system presented in this paper is based in two fundamental elements, an annotation system which is a modified version of ViPER Ground Truth tool and a tracking system optimized to perform video analysis in real time. The user supervisor monitors the automatic annotation process between these components through the ViPER-GT interface. The overall architecture of the proposed framework is called MViPER-GT and is illustrated in Fig. 1.

System works as follow. MViPER-GT tool sends raw frames from the video selected to the tracker when the analysis starts. Supervisor may update the track annotations during the analysis. These changes consist on create and delete tracks and adjust their size and position. Update track annotations can be done in real time if an error is detected, for instance, if a track is not created automatically by the tracking system or if a location prediction of a track is not properly done. MViPER-GT sends changes done by the supervisor to the tracking system. Information sent between subsystems is performed through a communication layer. This layer interacts in a bidirectional manner, transforming information to operate on each subsystem and enabling the communication. Bidirectional communication between VIPER-GT and the tracking system is performed one time per frame.

Tracking system is responsible for carrying out the establishment, updating and deleting tracks automatically. Once the tracker has received information concerning to MViPER-GT (a raw image and in some cases, also, user annotation updates), it realizes a complete tracking analysis, which includes segmentation, association, trajectory prediction, etc. User updates are taken into account during the analysis. Feature modifications may cause reaction in tracking system, changing the size or the location of the tracks. For instance, if a track trajectory is modified in the annotation tool, this will perform modifications in the calculated trajectory of the tracker, normally, trajectories followed at this time, will probably suffer alterations in the same sense of the new positioning. Thanks to the trajectory prediction algorithms housed in the tracking system, MViPER-GT receives non linear trajectories of each track or object in each frame. When the analysis is completed, tracking system predictions (updated feature tracks, size and position) are sent to the annotation tool. The annotation tool receives these predictions annotates them and starts a new cycle with the next frame.

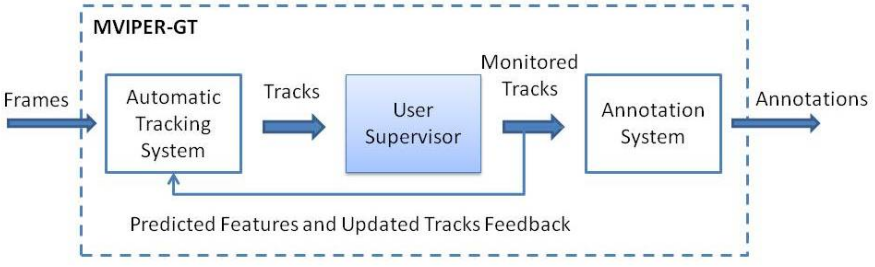

Fig. 1 Overall Architecture.

3.1 Annotation System: ViPER-GT and MViPER-GT

Create an annotation model was a tedious task, especially in the video domain, because it was required review sequences of frames with the similar content from frame to frame. Annotation system based on ViPER-GT makes it easier. ViPER-GT is a Java open source development project, supported by the ViPER API. When ViPER was thought, the first goal was the creation of a flexible ground truth format and second goal was to provide tools to easily create and share ground truth data [4]. Developed GUI could be used to record the requisite information in a single scan of the video content. For a given frame, users could select a cell representing a spatial attribute (point, bbox, obox or circle) [4].

The modified version of ViPER-GT presented in this paper, allows user to be merely a supervisor of the annotation task. The tool allows users to configure data generation and evaluation. Descriptors represent the data generation structure of each video. Structures of the objects are defined by the descriptors which can contain different kinds of attributes. Annotations describe the object state through its attributes. Attributes contain the feature values which can represent, for instance,

the name or the location of an object. MViPER-GT starts with two predefined descriptors, a static metadata structure which represents video information like the number of frames or the frame rate and a dynamic metadata structure which represents the object information. This structure has two attributes, the track identification number and the bounding box which represent the position and the size of each track. New descriptors may be created to this basic configuration, in order to separate different knowledge levels. Descriptors based, for example, on semantic information may be created before or during the analysis; however the annotations should be done manually.

ViPER offer some annotation possibilities like creation, deletion of tracks and modifications on the features of each track. As we seen before MViPER tool includes trajectory prediction algorithms. ViPER standard version includes a default linear interpolation utility which can also be used with MViPER. This linear interpolation utility fills in new intermediate values of spatial attributes between two separated frames.

Sometimes tracking system detects systematically a new track which the user does not want to annotate. ViPER propagation utility is the best way to treat this kind of situations. Propagation copies the current frame's value of selected object descriptors to all frames in the range of propagations [11]. This is especially helpful for spatial attributes that do not change much across frames. All attribute values of the chosen object descriptors are overridden with the values in the current frame [11]. Unwished labelling may be avoided, enabling propagation and disabling track annotations.

3.2 Tracking System

Architecture is based on a video chain with different modules that run in sequence, which correspond to the successive phases of the tracking process. The tracking system is composed by four modules: Foreground/Background Detection module shows when a pixel has moved and group them in blobs. Association module predicts the blobs positions, assign sets of blobs to tracks and finally update the tracks positions. Initialize or Delete module, create and delete tracks when have not assigned to any blob. Trajectory Generator module detects anomalous behaviours studying tracks trajectories. Algorithms belonging to each module are interchangeable in each run. Each module has a specific task to be implemented by all the algorithms that correspond to a certain module [10]. The input data for the pipeline is the image of current frame and the output data is the tracks position and size.

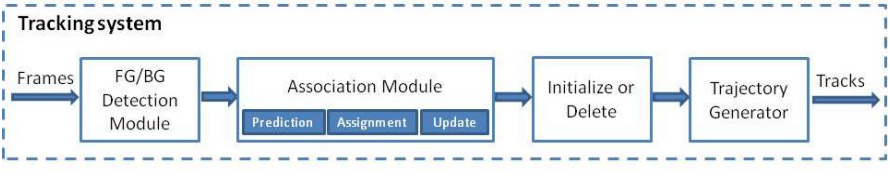

Fig. 2 Tracking system.

4 Experimental Results

We have implemented a prototype to get some experimental result. A JNI communication layer has been developed to achieve bidirectional interaction between ViPER-GT annotation system and OpenCV tracking system.

This system has been tested with the Computer Vision Based Analysis in Sport Environments (CVBASE) dataset [19]. Video features are 25 frames per second, 384x576 pixels of resolution and M-JPEG compression.

Selected video is a zenithal record of two players playing squash. They are in close proximity to each other, they are dressed similarly and are moving quickly, and there are constant crossings and occlusions between players, which make the video an interesting challenge to the quality measure of the system.

As we can see in this first image, the operation of the tracking system for this video is correct. Tracks are detected with quite accurately and the annotations are automatically done. However there is not still a complicated situation where the user has to intervene.

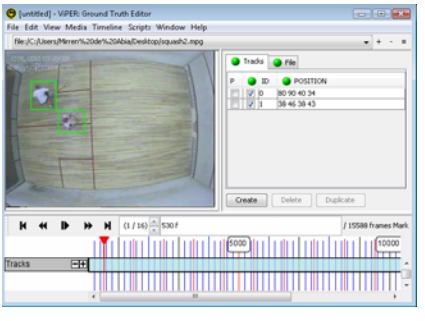

Fig. 3 Video analysis. Frame 530.

The two images below show the performance under critical circumstances of occlusion between tracks. The system has also an optimal behavior in such cases, therefore, it is not necessary, in this case, the intervention of the supervisor to make changes in the annotations.

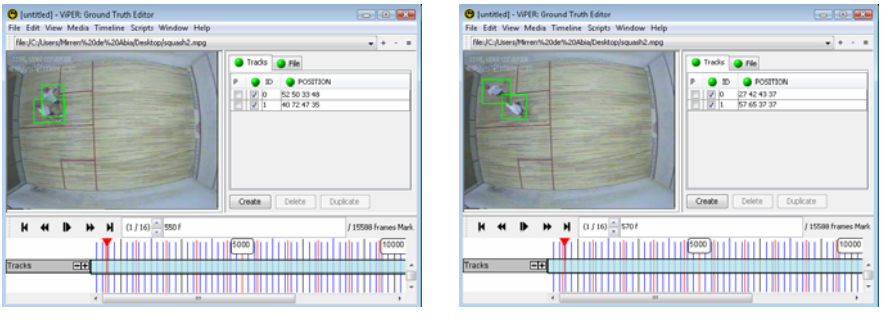

Fig. 4 Video analysis. **4A** Frame 550. **4B** Frame 570.

In general this modified version of ViPER improves notably annotation times. In many cases it is not necessary to realize changes from one frame to other. In some cases only it is necessary to do small and simple modifications in a fully annotated frame.

Ground truths are stored in XML files as sets of descriptor records. Each descriptor annotates an associated range of frames by instantiating a set of attributes for that range [4]. This is a sample code which represents the track position in a range of frames. Positions are denoted by bounding boxes.

```
<object framespan="-2147483648:2147483646" id="2" name="Tracks">
    <attribute name="POSITION">
        <data:bbox framespan="146:148" height="11"
            width="16" x="26" y="72"/>
        <data:bbox framespan="301:303" height="10"
            width="15" x="26" y="72"/>
    </attribute>
</object>
```

Fig. 5 Sample code.

Manually semantic annotation it is possible, thanks to the ViPER Schema Editor utility. Adding a new descriptor and attributes it is feasible to carry out a semantic description about what is happening in the video. This images show how users can annotate the name of the players and when it is producing a cross between them.

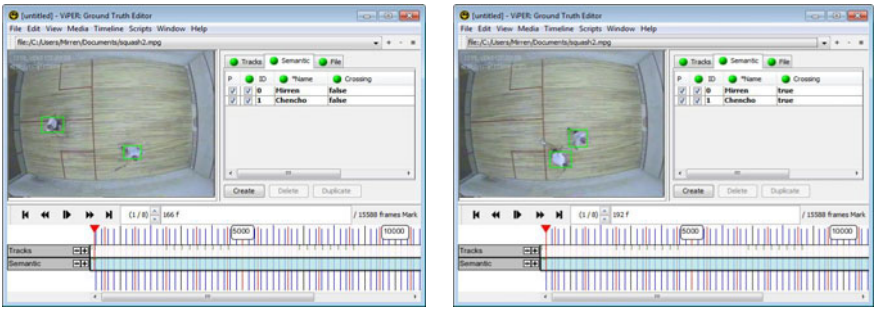

Fig. 6 Semantic analysis. **6A** Frame 166. **6B** Frame 192.

5 Conclusion

We have presented a new annotation tool for interactive *ground-tuth* generation. This system integrates a tracking module for a multi-level automatic and supervised labeling. In general MViPER-GT improves notably annotation times. In many cases it is not necessary to realize changes from one frame to other, and only in some cases it is necessary to do small and simple modifications in a fully annotated frame.

Future works will be addressed to the configuration of the tracking system through a XML file, the integration of a context-based module to introduce

semi-automatic annotation at scene and object level and the capability to develop automatic annotations at the semantic level based on the behavior of the tracked elements.

References

1. Snoek, C.G.M., Worring, M.: Multimodal Video Indexing: A Review of the State-of-the-Art. Multimedia Tools and Applications 25(1), 5–35 (2004)
2. Bloehdorn, S., Petridis, K., Saathoff, K., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, Y., Staab, S., Strintzis, M.G.: Semantic Annotation of Images and Videos for Multimedia Analysis. In: Gómez-Pérez, A., Euzenat, J. (eds.) ESWC 2005. LNCS, vol. 3532, pp. 592–607. Springer, Heidelberg (2005)
3. Butler, M., Zapart, T., Li, R.: Video Annotation – Improving Assessment of Transient Educational Events. In: Proceedings of the 2006 Informing Science and IT Education Joint Conference (2006)
4. Doermann, D., Mihalcik, D.: Tools and Techniques for Video Performance Evaluation. In: 15th International Conference on Pattern Recognition, vol. 4, p. 4167 (2000)
5. Panagi, P., Dasiopoulou, S., Papadopoulos, G.T., Kompatsiaris, I., Strintzis, M.G.: A Genetic Algorithm Approach Ontology-Driven Semantic Image Analysis. In: IET International Conference on Visual Information Engineering, pp. 132–137 (2006)
6. Black, J., Ellis, T., Rosin, P.: A Novel Method for Video Tracking Performance Evaluation. In: Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (2003)
7. Assfalg, J., Bertini, M., Colombo, C., Del Bimbo, A.: Semantic Annotation of Sports Videos. IEEE Multimedia Magazine 9(2), 52–60 (2002)
8. Kender, J.R., Naphade, M.R.: Visual Concepts for News Story Tracking: Analyzing and Exploiting the NIST TRECVID Video Annotation Experiment. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 1174–1181 (2005)
9. D'Orazio, T., Leo, M., Mosca, N., Spagnolo, P., Mazzeo, P.L.: A Semi-Automatic System for Ground Truth Generation of Soccer Video Sequences. In: 2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance, pp. 559–564 (2009)
10. Sánchez, A.M., Patricio, M.A., García, J., Molina, J.M.: A Context Model and Reasoning System to Improve Object Tracking in Complex Scenarios. Expert Systems with Applications 36, 10995–11005 (2009)
11. Language and Media Processing Laboratory. The Video Performance Evaluation Resource, `http://viper-toolkit.sourceforge.net`
12. Surveillance Performance EValuation Initiative (SPEVI), `http://www.elec.qmul.ac.uk/staffinfo/andrea/spevi.html`
13. A chroma-based Video Segmentation Ground-truth, `http://www-vpu.ii.uam.es/CVSG/`
14. OTCBVS Benchmark Dataset Collection, `http://www.cse.ohio-state.edu/otcbvs-bench/`
15. Video Surveillance Online Repository (VISOR), `http://imagelab.ing.unimore.it/visor/video_categories.asp`
16. ETISEO Video undestanding Evaluation, `http://www-sop.inria.fr/orion/ETISEO/`
17. CANDELA project, `http://www.multitel.be/~va/candela/`
18. Computational Vision Group, `http://www.cvg.rdg.ac.uk/`
19. CVBASE dataset, `http://vision.fe.uni-lj.si/cvbase06/downloads.html`